



(12) **United States Patent**
Bernbo et al.

(10) **Patent No.:** **US 9,305,012 B2**
(45) **Date of Patent:** ***Apr. 5, 2016**

(54) **METHOD FOR DATA MAINTENANCE**

(56) **References Cited**

(71) Applicant: **COMPUVERDE AB**, Karlskrona (SE)

U.S. PATENT DOCUMENTS

(72) Inventors: **Stefan Bernbo**, Karlskrona (SE);
Christian Melander, Rodeby (SE);
Rodger Persson, Karlskrona (SE);
Gustav Petersson, Sturko (SE)

3,707,707	A	12/1972	Spencer et al.
5,787,247	A	7/1998	Norin et al.
6,003,065	A	12/1999	Yan et al.
6,021,118	A	2/2000	Houck et al.
6,055,543	A	4/2000	Christensen et al.

(Continued)

(73) Assignee: **COMPUVERDE AB** (SE)

FOREIGN PATENT DOCUMENTS

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

CN	1726454	A	1/2006
EP	0774723	A3	7/1998

(Continued)

This patent is subject to a terminal disclaimer.

OTHER PUBLICATIONS

(21) Appl. No.: **14/098,021**

Stephen, Multicast Routing in Datagram Internetworks and Extended LANs, 1990.*

(22) Filed: **Dec. 5, 2013**

(Continued)

(62) **Prior Publication Data**

US 2014/0095559 A1 Apr. 3, 2014

Related U.S. Application Data

(63) Continuation of application No. 13/224,415, filed on Sep. 2, 2011, now Pat. No. 8,645,978.

Primary Examiner — Lechi Truong

(74) *Attorney, Agent, or Firm* — Condo Roccia Koptiw LLP

(51) **Int. Cl.**

G06F 3/00	(2006.01)
G06F 17/30	(2006.01)
G06F 9/54	(2006.01)
H04L 12/24	(2006.01)

(57)

ABSTRACT

A method for data storage implemented in a data storage system is disclosed. Data storage nodes may be interconnected by means of a communications network. The method may include sending a request for a first data item to a plurality of storage nodes. The first data item may include a reference to a second data item stored in the storage system. The method may include receiving the first data item from at least one storage node, and sending a request for the second data item to the plurality of storage nodes based on the reference included in the first data item.

(52) **U.S. Cl.**

CPC **G06F 17/30194** (2013.01); **G06F 9/54** (2013.01); **G06F 17/30206** (2013.01); **G06F 17/30212** (2013.01); **H04L 41/0213** (2013.01)

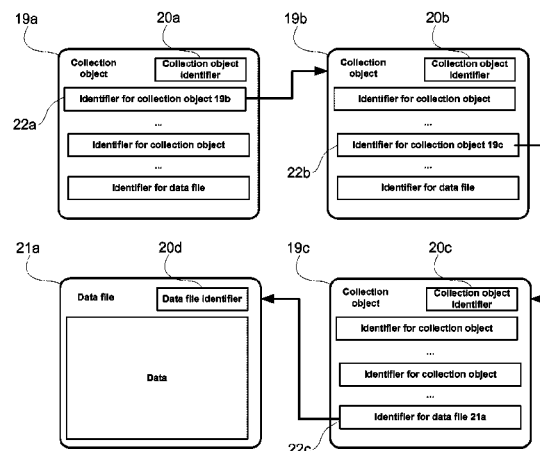
(58) **Field of Classification Search**

CPC G06F 9/45; G06F 17/30; G06F 9/54; H04L 41/2013

USPC 719/310, 318; 709/223

See application file for complete search history.

18 Claims, 9 Drawing Sheets



(56)

References Cited

U.S. PATENT DOCUMENTS

6,389,432	B1	5/2002	Pothapragada et al.
6,470,420	B1	10/2002	Hospodor
6,782,389	B1	8/2004	Chrin et al.
6,925,737	B2	8/2005	Bolduan et al.
6,985,956	B2	1/2006	Luke et al.
7,039,661	B1	5/2006	Ranade et al.
7,200,664	B2	4/2007	Hayden
7,206,836	B2	4/2007	Dinker et al.
7,266,556	B1	9/2007	Coates
7,320,088	B1	1/2008	Gawali et al.
7,340,510	B1	3/2008	Liskov et al.
7,352,765	B2	4/2008	Dai et al.
7,406,484	B1	7/2008	Srinivasan et al.
7,487,305	B2	2/2009	Hill et al.
7,503,052	B2	3/2009	Castro et al.
7,546,486	B2	6/2009	Slik et al.
7,568,069	B2	7/2009	Jantz et al.
7,574,488	B2	8/2009	Matsubara
7,590,672	B2	9/2009	Slik et al.
7,593,966	B2	9/2009	Therrien et al.
7,624,155	B1	11/2009	Nordin et al.
7,624,158	B2	11/2009	Slik et al.
7,631,023	B1	12/2009	Kaiser et al.
7,631,045	B2	12/2009	Boerries et al.
7,631,313	B2	12/2009	Mayhew et al.
7,634,453	B1	12/2009	Bakke et al.
7,647,329	B1	1/2010	Fischman et al.
7,694,086	B1	4/2010	Kulkarni et al.
7,769,711	B2	8/2010	Srinivasan et al.
7,778,972	B1	8/2010	Cormie et al.
7,822,766	B2	10/2010	Arndt et al.
7,840,992	B1	11/2010	Dufrene et al.
7,873,650	B1	1/2011	Chapman et al.
7,885,982	B2	2/2011	Wight et al.
8,060,598	B1	11/2011	Cook et al.
8,073,881	B1	12/2011	Georgiev
8,190,561	B1	5/2012	Poole et al.
8,255,430	B2	8/2012	Dutton et al.
8,296,398	B2*	10/2012	Lacapra et al. 709/219
8,401,997	B1	3/2013	Tawri et al.
8,417,828	B2	4/2013	Ma et al.
8,443,062	B2	5/2013	Voutilainen et al.
8,561,115	B2	10/2013	Hattori et al.
8,577,957	B2	11/2013	Behar et al.
8,707,091	B2*	4/2014	Gladwin et al. 714/6.24
2001/0034812	A1	10/2001	Ignatius et al.
2001/0047400	A1	11/2001	Coates et al.
2002/0042693	A1	4/2002	Kampe et al.
2002/0073086	A1	6/2002	Thompson et al.
2002/0103888	A1	8/2002	Janz et al.
2002/0114341	A1	8/2002	Sutherland et al.
2002/0145786	A1	10/2002	Chang et al.
2003/0026254	A1	2/2003	Sim
2003/0120654	A1	6/2003	Edlund et al.
2003/0126122	A1	7/2003	Bosley et al.
2003/0154238	A1	8/2003	Murphy et al.
2003/0172089	A1	9/2003	Douceur et al.
2003/0177261	A1	9/2003	Sekiguchi et al.
2004/0059805	A1	3/2004	Dinker et al.
2004/0064729	A1	4/2004	Yellepeddy
2004/0078466	A1	4/2004	Coates et al.
2004/0088297	A1	5/2004	Coates et al.
2004/0111730	A1	6/2004	Apte
2004/0243675	A1	12/2004	Taoyama et al.
2004/0260775	A1	12/2004	Fedele
2005/0010618	A1	1/2005	Hayden
2005/0015431	A1	1/2005	Cherkasova
2005/0015461	A1	1/2005	Richard et al.
2005/0038990	A1	2/2005	Sasakura et al.
2005/0044092	A1	2/2005	Adya et al.
2005/0055418	A1	3/2005	Blanc et al.
2005/0177550	A1	8/2005	Jacobs et al.
2005/0193245	A1	9/2005	Hayden et al.
2005/0204042	A1	9/2005	Banerjee et al.
2005/0246393	A1	11/2005	Coates et al.
2005/0256894	A1	11/2005	Talanis et al.
2005/0278552	A1	12/2005	Delisle et al.
2005/0283649	A1	12/2005	Turner et al.
2006/0031230	A1	2/2006	Kumar
2006/0031439	A1	2/2006	Saffre
2006/0080574	A1	4/2006	Saito et al.
2006/0090045	A1	4/2006	Bartlett et al.
2006/0090095	A1	4/2006	Massa et al.
2006/0112154	A1	5/2006	Douceur et al.
2006/0218203	A1	9/2006	Yamato et al.
2007/0022087	A1	1/2007	Bahar et al.
2007/0022121	A1	1/2007	Bahar et al.
2007/0022122	A1	1/2007	Bahar et al.
2007/0022129	A1	1/2007	Bahar et al.
2007/0055703	A1*	3/2007	Zimran et al. 707/200
2007/0088703	A1	4/2007	Kasiolas et al.
2007/0094269	A1	4/2007	Mikesell et al.
2007/0094354	A1	4/2007	Soltis
2007/0189153	A1	8/2007	Mason
2007/0198467	A1	8/2007	Wiser et al.
2007/0220320	A1	9/2007	Sen et al.
2007/0276838	A1	11/2007	Abushanab et al.
2007/0288494	A1	12/2007	Chrin et al.
2007/0288533	A1	12/2007	Srivastava et al.
2007/0288638	A1	12/2007	Vuong et al.
2008/0005199	A1	1/2008	Chen et al.
2008/0043634	A1	2/2008	Wang et al.
2008/0077635	A1	3/2008	Sporny et al.
2008/0104218	A1	5/2008	Liang et al.
2008/0109830	A1	5/2008	Giotzbach et al.
2008/0168157	A1	7/2008	Marchand
2008/0171556	A1	7/2008	Carter
2008/0172478	A1	7/2008	Kiyohara et al.
2008/0198752	A1	8/2008	Fan et al.
2008/0235321	A1	9/2008	Matsuo
2008/0244674	A1	10/2008	Hayashi et al.
2008/0270822	A1	10/2008	Fan et al.
2009/0043922	A1	2/2009	Crowther
2009/0083810	A1	3/2009	Hattori et al.
2009/0132543	A1	5/2009	Chatley et al.
2009/0172211	A1	7/2009	Perry et al.
2009/0172307	A1	7/2009	Perry et al.
2009/0228669	A1	9/2009	Slesarev et al.
2009/0271412	A1	10/2009	Lacapra et al.
2009/0287842	A1	11/2009	Plamondon
2010/0115078	A1	5/2010	Ishikawa et al.
2010/0161138	A1	6/2010	Lange et al.
2010/0169391	A1	7/2010	Baptist et al.
2010/0169415	A1	7/2010	Leggette et al.
2010/0185693	A1	7/2010	Murty et al.
2010/0198888	A1	8/2010	Blomstedt et al.
2010/0198889	A1	8/2010	Byers et al.
2010/0223262	A1	9/2010	Krylov et al.
2010/0303071	A1	12/2010	Kotalwar et al.
2011/0055353	A1	3/2011	Tucker et al.
2011/0072206	A1	3/2011	Ross et al.
2011/0125814	A1	5/2011	Slik et al.
2011/0252204	A1	10/2011	Coon et al.
2012/0180070	A1*	7/2012	Pafumi et al. 719/313
2012/0331021	A1	12/2012	Lord
2013/0060884	A1	3/2013	Bernbo et al.
2013/0103851	A1	4/2013	Umeki et al.
2013/0254314	A1	9/2013	Chow et al.

FOREIGN PATENT DOCUMENTS

EP	0934568	B1	6/2003
EP	1521189	A2	4/2005
EP	1578088	A2	9/2005
EP	1669850	A1	6/2006
EP	1798934	A1	6/2007
EP	2031513	A2	3/2009
JP	6-348527	A	12/1994
JP	11-249874	A	9/1999
JP	2000-322292	A	11/2000
JP	2003-030012	A	1/2003
JP	2003-223286	A	8/2003
JP	2003-248607	A	9/2003
JP	2003-271316	A	9/2003

(56)

References Cited

FOREIGN PATENT DOCUMENTS

JP	2004-005491	A	1/2004
JP	2007-058275	A	3/2007
JP	2008-250767	A	10/2008
JP	2009-259007	A	11/2009
WO	WO 99/38093	A1	7/1999
WO	WO 01/18633	A1	3/2001
WO	WO 02/35359	A2	5/2002
WO	WO 02/44835	A2	6/2002
WO	WO 2004/053677	A2	6/2004
WO	WO 2006/124911	A2	11/2006
WO	WO 2007/014296	A2	2/2007
WO	WO 2007/115317	A2	10/2007
WO	WO 2007/134918	A1	11/2007
WO	WO 2008/069811	A1	6/2008
WO	WO 2008/102195	A1	8/2008
WO	WO 2009/048726	A1	4/2009
WO	WO 2010/046393	A2	4/2010
WO	WO 2010/080533	A2	7/2010
WO	WO 2011/131717	A1	10/2011

OTHER PUBLICATIONS

Wikipedia, "FastTrack", Available online at: <http://de.wikipedia.org/w/index.php?title=FastTrack&01did=83614953>, Jan. 8, 2011, pp. 1-2.

Wikipedia, "Load Balancing (Computing)", Available online at http://en.wikipedia.org/w/index.php?title=Load_balancing_%28computing%29&01did=446655159, Aug. 25, 2011, pp. 1-7.

SAP Library, "Queues for Prioritized Message Processing", SAP Exchange Infrastructure, Available online at http://help.sap.com/saphelp_nw04/helpdata/en/04/827440c36ed562e10000000a155106/content.htm, Feb. 6, 2009, pp. 1-2.

Squid-Cache.Org, "Squid Configuration Directive Reply_Body_Max_Size", Available online at < squid-cache.org/Doc/config/reply_body_max_size/>, Dec. 21, 2008, pp. 1-2.

Trustwave, "How Do I Block Large Files by Content Size Before Download?", Available online at < trustwave.com/support/kb/article.aspx?id=13166>, Mar. 23, 2009, pp. 1-2.

Hewlett-Packard Development Company L. P., "HP Volume Shadowing for OpenVMS", OpenVMS Alpha 7.3-2, Sep. 2003, 162 pages.

Kronenberg et al., "VAXclusters: A Closely-Coupled Distributed System", ACM Transactions on Computer Systems, vol. 4, No. 2, May 1986, pp. 130-146.

Parris, Keith, "Using OpenVMS Clusters for Disaster Tolerance", System/Software Engineer, HP Services—Systems Engineering, 2003, 27 pages.

Tang et al., "An Efficient Data Location Protocol for Self-Organizing Storage Clusters", Supercomputing, ACM/IEEE Conference, Phoenix, AZ, USA, Nov. 15-21, 2003, 13 pages.

Weatherspoon et al., "Antiquity: Exploiting a Secure Log for Wide-Area Distributed Storage", Proceedings of the EuroSys Conference, ACM 2007, Lisbon, Portugal, Mar. 21-23, 2007, pp. 371-384.

Zhang et al., "Brushwood: Distributed Trees in Peer-to-Peer Systems", Peer-to-Peer Systems IV Lecture Notes in Computer Science vol. 3640, 2005, pp. 47-57.

Katsurashima et al., "NAS Switch: a Novel CIFS Server Virtualization", Proceedings. 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies, Apr. 7-10, 2003, pp. 82-86.

Suryanarayanan et al., "Performance Evaluation of New Methods of Automatic Redirection for Load Balancing of Apache Servers Distributed in the Internet", Proceedings. 25th Annual IEEE Conference on Local Computer Networks, Nov. 8-10, 2000, pp. 644-651.

* cited by examiner

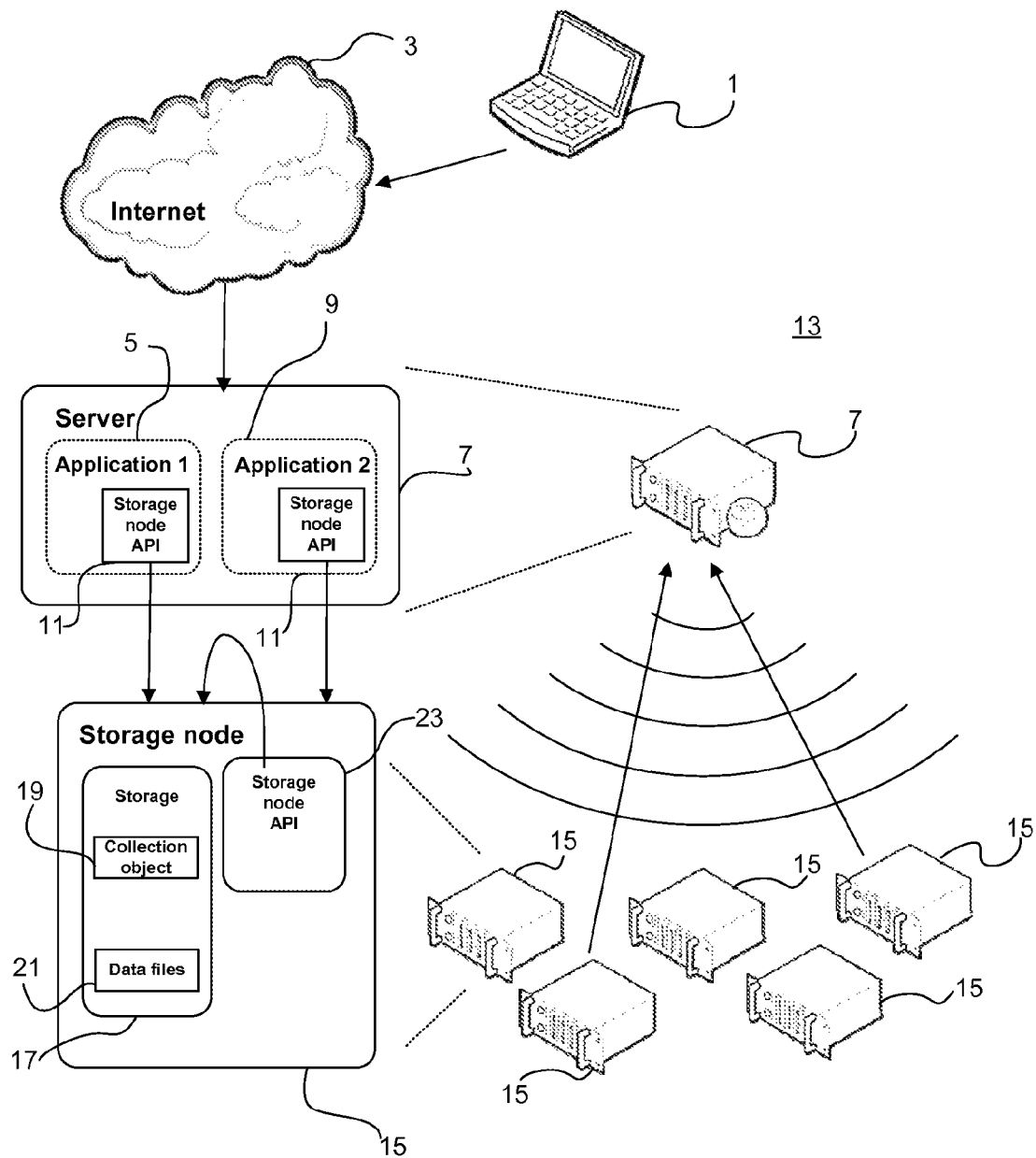


Fig 1

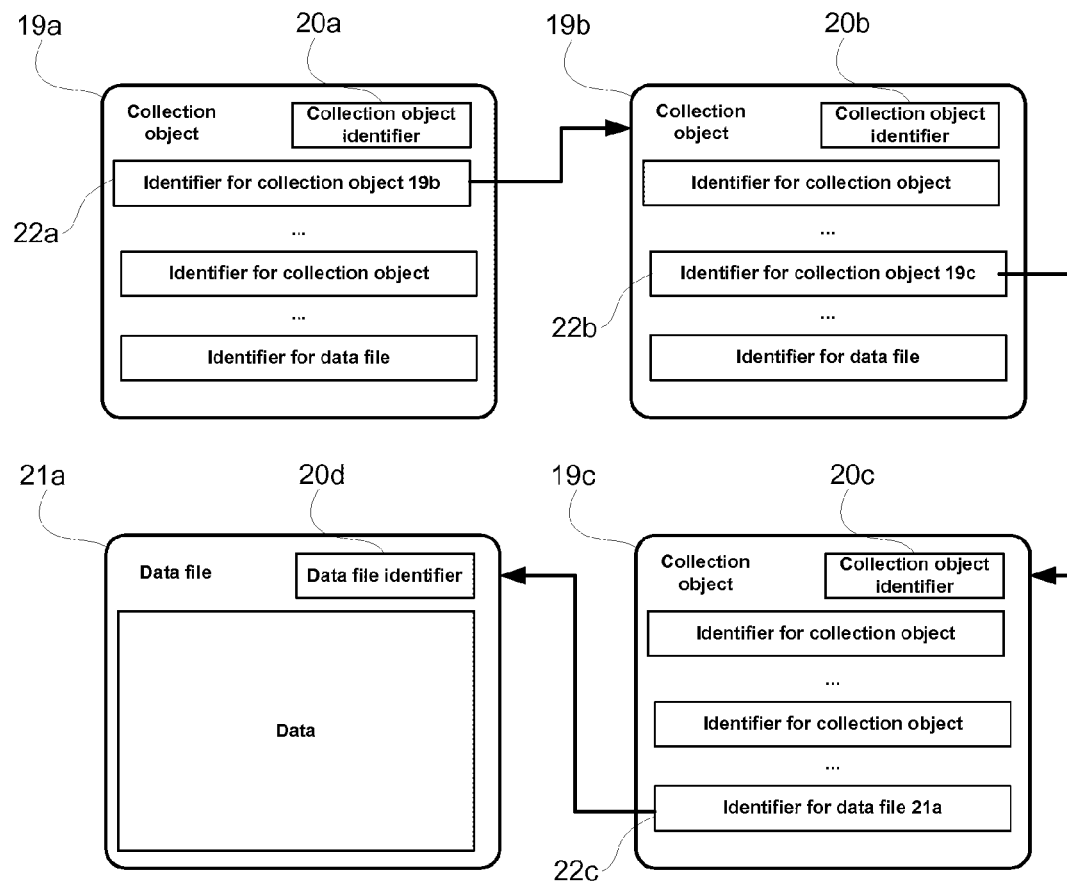


Fig 2

20a-d

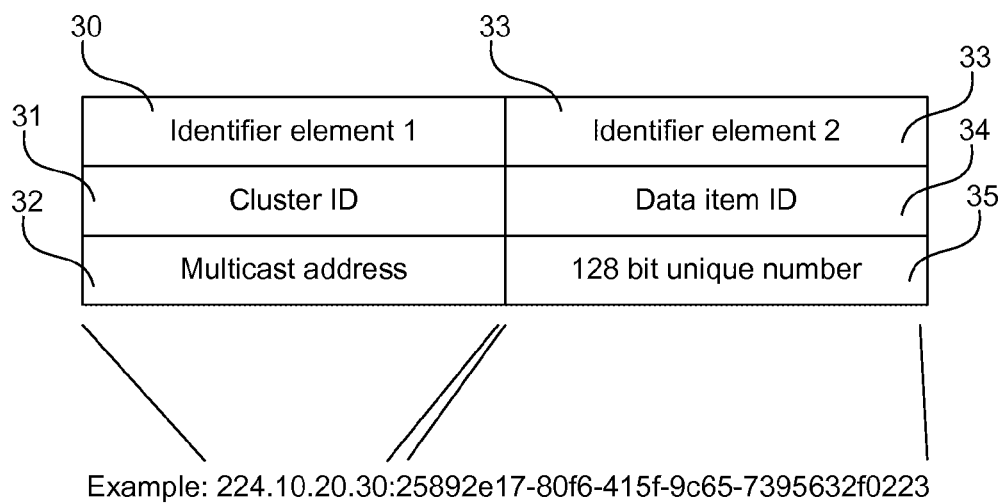


Fig 3

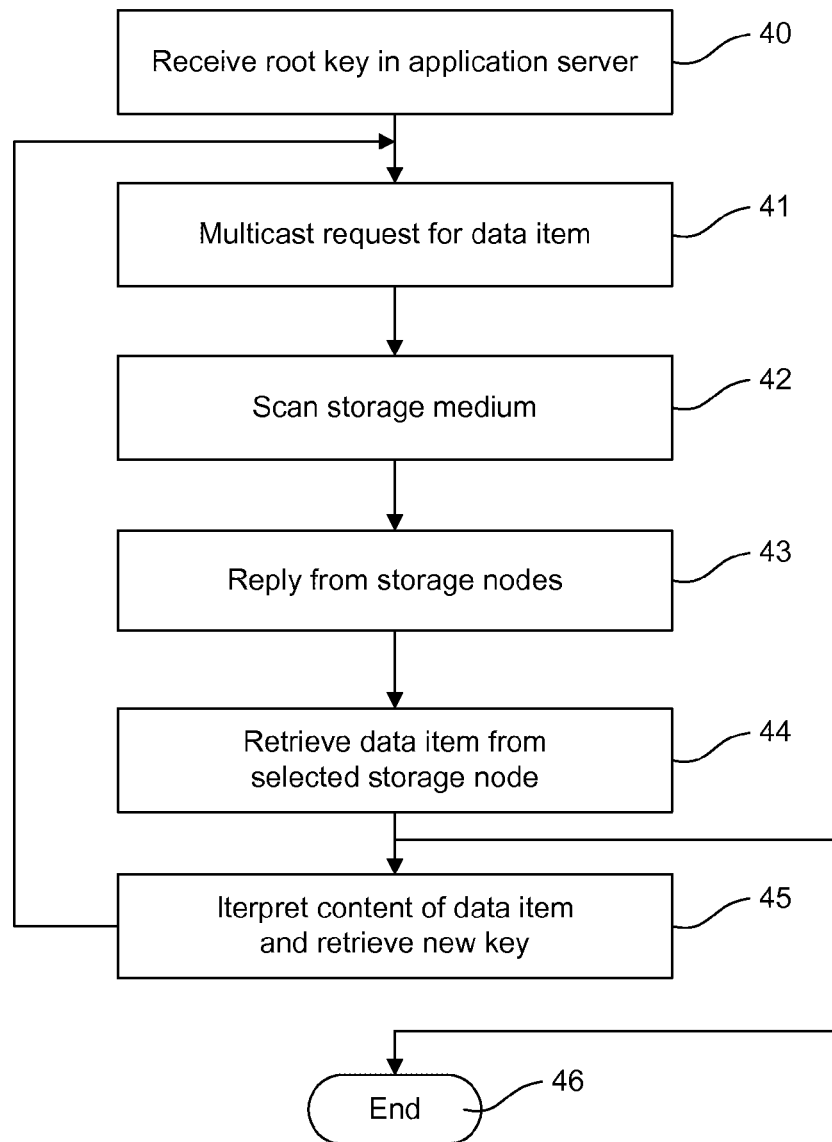


Fig 4

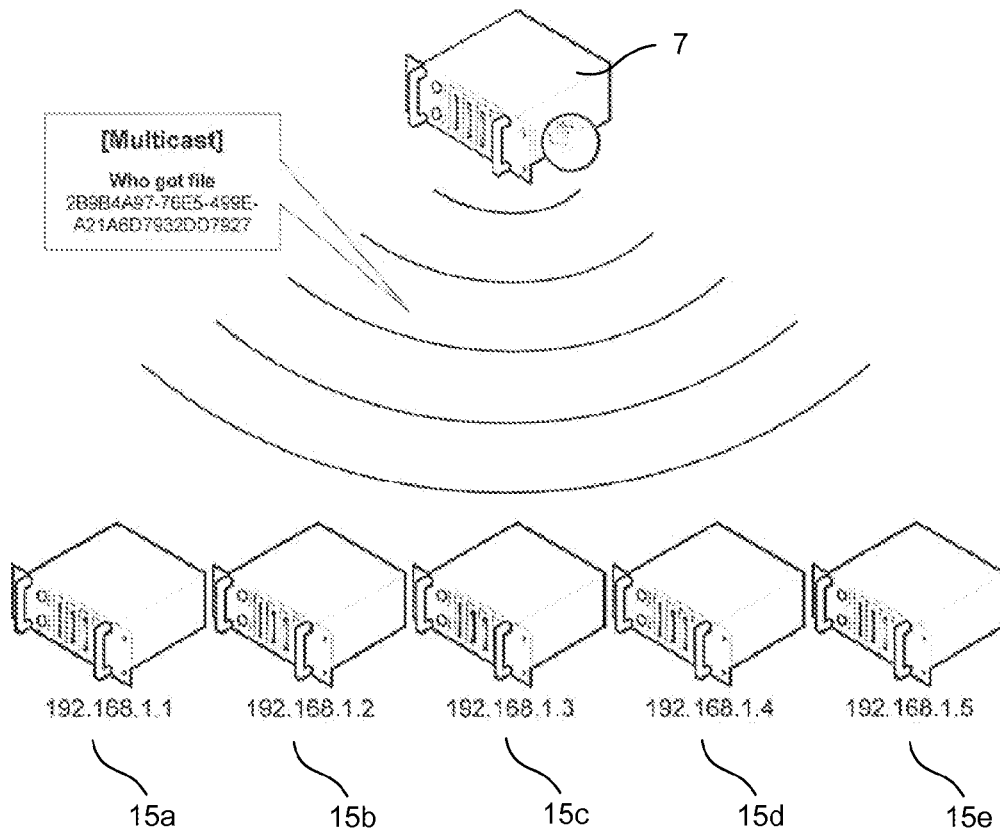


Fig 5a

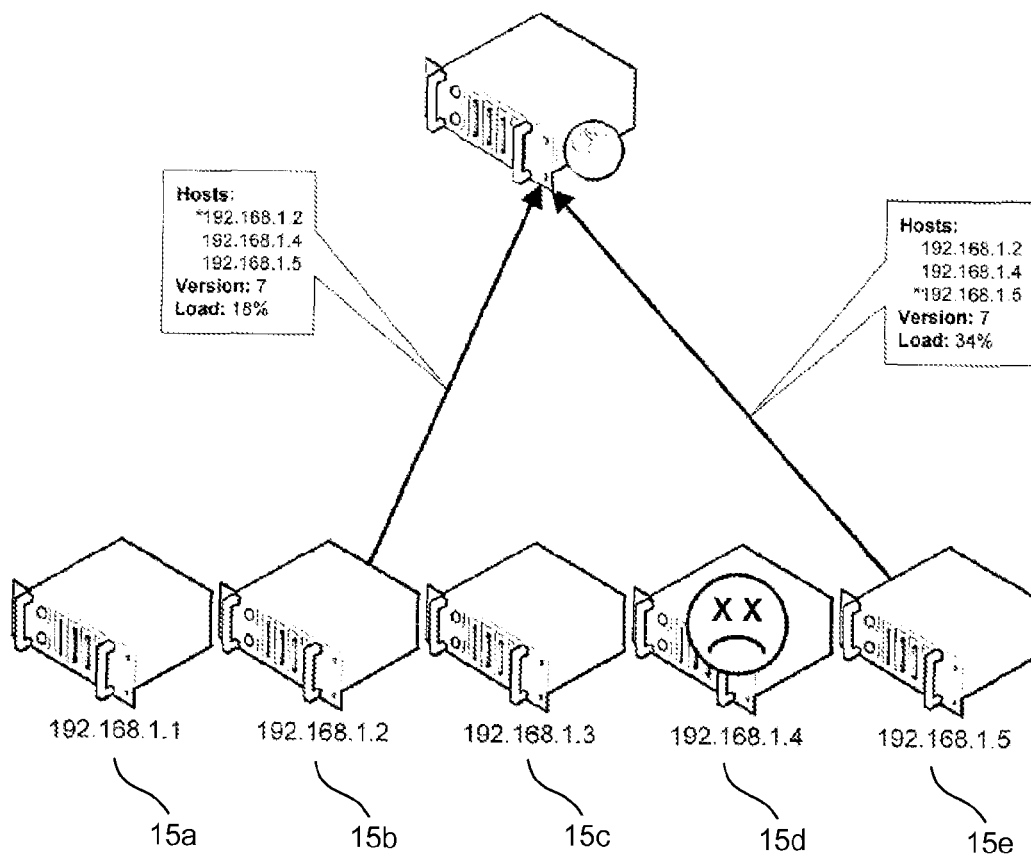


Fig 5b

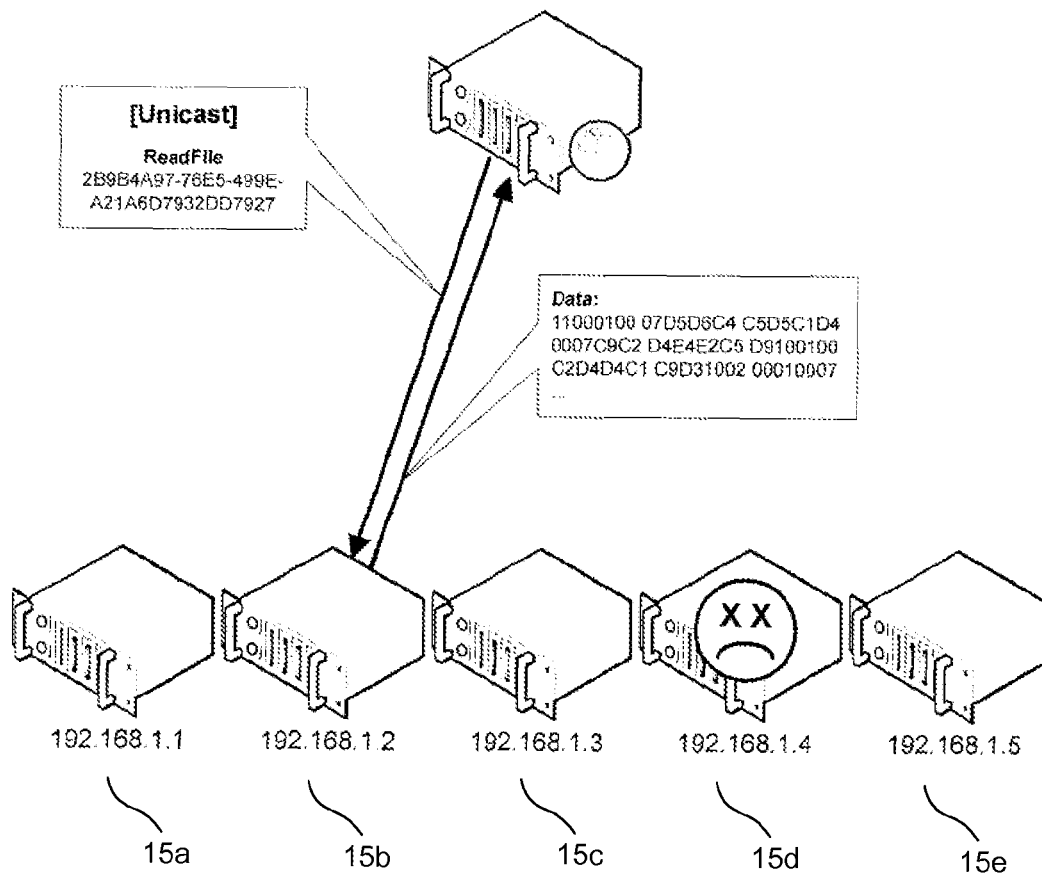


Fig 5c

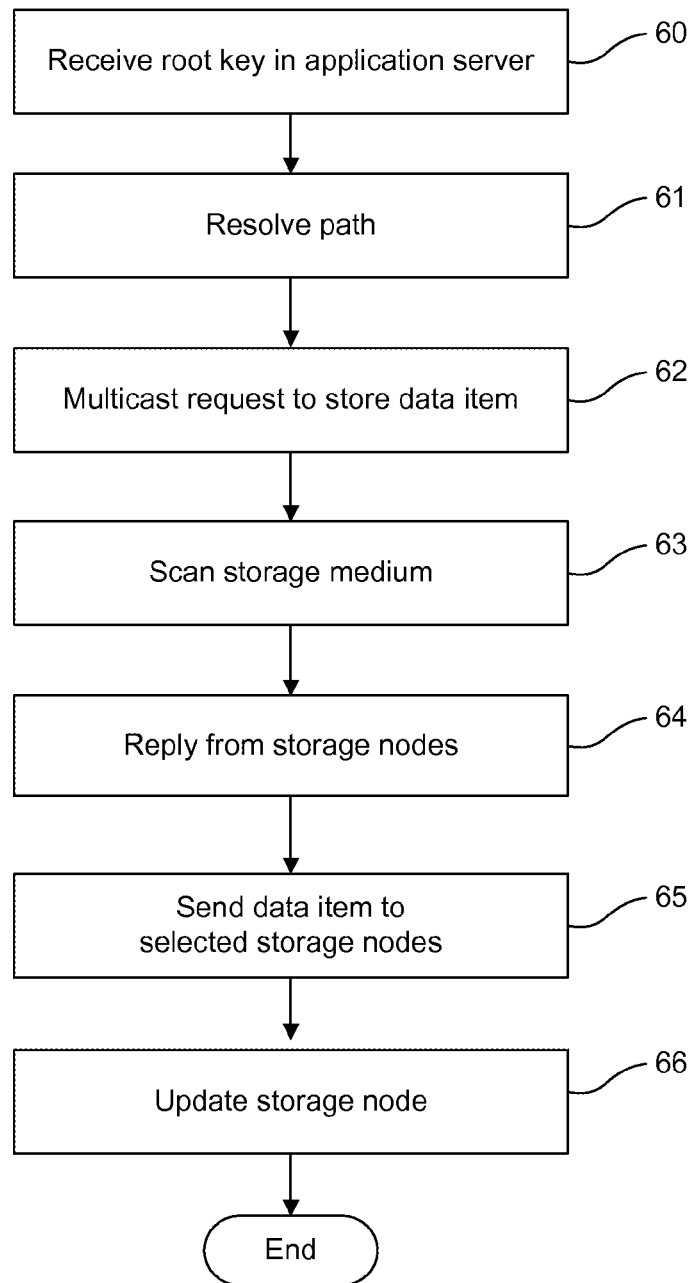
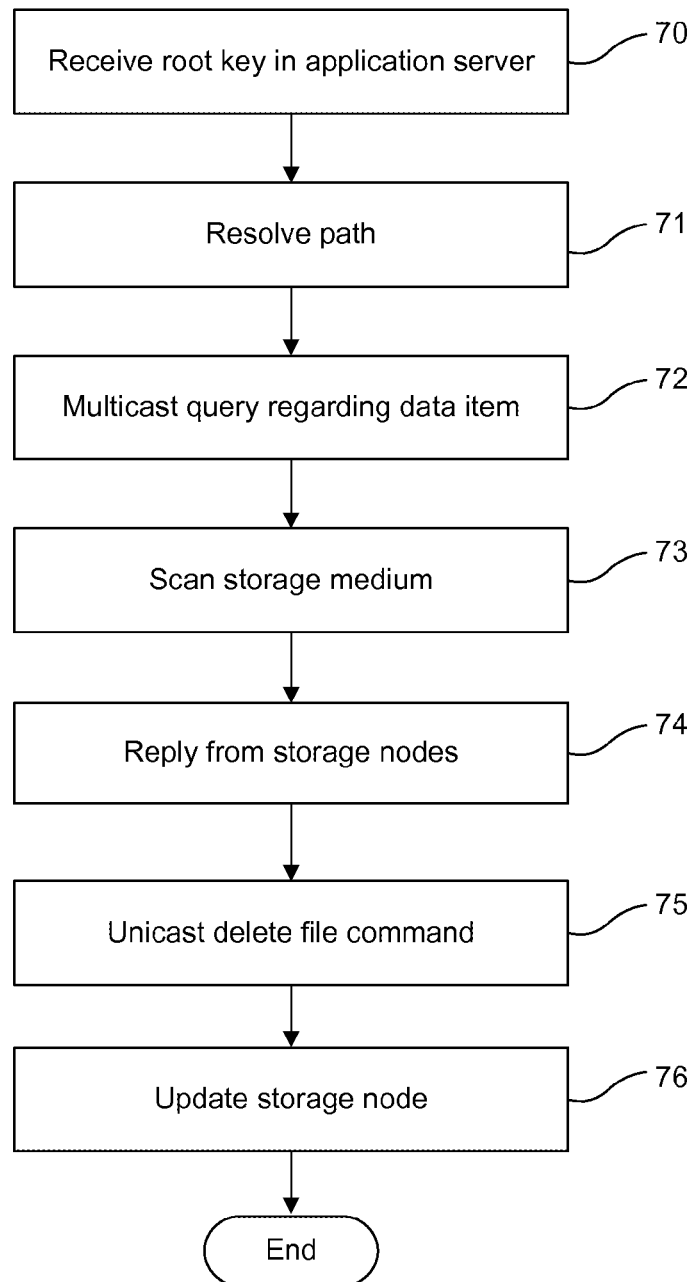


Fig 6

**Fig 7**

1

METHOD FOR DATA MAINTENANCE**CROSS REFERENCE TO RELATED APPLICATIONS**

This application is a continuation of U.S. patent application Ser. No. 13/224,415, filed Sep. 2, 2011, which issued as U.S. Pat. No. 8,645,978 on Feb. 4, 2014, the contents of which are hereby incorporated by reference herein.

This application includes subject matter that may be related to subject matter included in the following applications: U.S. patent application Ser. No. 13/224,393 entitled "A Method And Device For Maintaining Data In A Data Storage System Comprising A Plurality Of Data Storage Nodes," filed Sep. 2, 2011, which issued as U.S. Pat. No. 8,650,365 on Feb. 11, 2014; U.S. patent application Ser. No. 13/224,404 entitled "Method And Device For Writing Data In A Data Storage System Comprising A Plurality Of Data Storage Nodes," filed Sep. 2, 2011, which issued as U.S. Pat. No. 9,021,053 on Apr. 28, 2015; U.S. patent application Ser. No. 13/224,424 entitled "Method For Handling Requests In A Storage System And A Storage Node For A Storage System," filed Sep. 2, 2011; U.S. patent application Ser. No. 13/224,433 entitled "Method For Data Retrieval From A Distributed Data Storage System," filed Sep. 2, 2011, which issued as U.S. Pat. No. 8,769,138 on Jul. 1, 2014; and U.S. patent application Ser. No. 13/224,446 entitled "Method For Updating Data In A Distributed Data Storage System," filed Sep. 2, 2011, which issued as U.S. Pat. No. 8,997,124 on Mar. 31, 2015.

TECHNICAL FIELD

The present disclosure relates to an apparatus and method for accessing, writing and deleting data in a data storage system comprising a plurality of data storage nodes, the methods may be employed in a server and/or in a storage node in the data storage system. The disclosure further relates to storage nodes or servers that may be capable of carrying out such methods.

BACKGROUND

Such a method is disclosed in US Patent Publication No. 2005/0246393 A1. This method is disclosed for a system that may use a plurality of storage centers at geographically disparate locations. Distributed object storage managers may be included to maintain information regarding stored data. One problem associated with such a system is how to accomplish simple and yet robust and reliable maintenance of data.

SUMMARY OF THE INVENTION

A method to maintain data in a data storage system comprising data storage nodes interconnected by means of a communications network is disclosed. The method may include sending a request for a first data item to a plurality of storage nodes. The first data item may include a reference to a second data item stored in the storage system. The method may also include receiving the first data item from at least one storage node, and/or sending a request for the second data item to the plurality of storage nodes, for example based on the reference included in the first data item.

In an embodiment, a virtual directory structure may be implemented in a storage system where the files are stored in an unstructured manner.

The request for the first and second data entities may be sent by multicast.

2

For example, by utilizing multicast many storage nodes may be easily accessed.

The first and second data items may be identified by a first and second unique key.

5 In an embodiment, the files in the storage system may be accessed directly irrespective of their location in the system.

The first and second unique key may include a cluster address pointing out a subset of said storage nodes within the system, and/or a data item identifier identifying a data item within the subset of storage nodes. A reference to a data item may include the unique key.

In an embodiment, the method may be implemented in large storage systems, for example a storage system which include hundreds or thousands of storage nodes.

15 The method may comprise sending the first and second request from an application programming interface, API.

For example, by utilizing a common API to access the storage nodes, the method may be easily implemented on many different platforms.

20 The API may be implemented on a server in communication with the storage nodes.

In an embodiment, the method may be implemented on a dedicated device provided e.g. by a third party which may not be responsible for the maintenance of the storage nodes.

25 The API may be implemented in a storage node.

In an example embodiment, implementing the API in a storage node may allow the number of access points into the storage system may be increased.

30 The method may comprise the API retrieving the unique key, identifying the second data item, from the received first data item.

For example, the unique identifier for the second data item may be easily retrievable if an indication of the unique identifier is included in the first data item.

35 The method may comprise receiving the key, identifying the first data item, in the API.

In an embodiment, one or more directory structures may be implemented simultaneously.

40 The second data item may include a reference to third data item.

For example, in an embodiment directory structures with multiple levels may be implemented.

The second data item may include payload data, such as an image.

45 For example, data files with payload data may be part of the directory structure (e.g. stored in sub folders).

The first data item may be sent by unicast.

For example, by utilizing unicast, the data item may be transferred in a bandwidth-effective manner.

50 According to an embodiment, a method for maintaining data in a data storage system including data storage nodes interconnected by means of a communications network may be implemented in a server and/or a data storage node. The method may include storing a first data item in at least one storage node. The method may also include updating a second data item, stored in at least one storage node. For example, the second data item may be updated by adding a reference to the first data item in the second data item. Updating the second data item may include sending a request to at least one storage node that stores a copy of the second data item. The request may instruct and/or request that the at least one storage node add a reference to the first data item to the second data item.

55 In an embodiment, new items may easily be added to the directory structure, for example by adding references to the new items to other items in the directory structure.

65 According to an embodiment, a method for maintaining data may be implemented a server or data storage node

3

included in a data storage system. The data storage nodes may be interconnected by means of a communications network. The method may include deleting a first data item stored in at least one storage node. The method may also include updating a second data item, stored in at least one storage node, by deleting a reference to the second data item in the second data item.

In an example embodiment, items in the directory structure may easily be deleted, for example by deleting references to the items.

According an embodiment, a data storage system may include data storage nodes interconnected by means of a communications network. A server or node may include an application programming interface, API, and may be configured to send a request for a first data item to a plurality of storage nodes. The first data item may include a reference to a second data item stored in the storage system. At least one storage node may be configured to send the first data item to the API. The API and/or storage node or sever may be further configured to send a request for a second data item to the plurality of storage nodes based on the reference included in the first data item.

For example, a virtual directory structure may be implemented in a storage system where the files are stored in an unstructured manner.

Other objectives, features and advantages of the disclosed embodiments may appear from the following detailed disclosure, from the attached claims as well as from the drawings.

Generally, all terms used in the claims are to be interpreted according to their ordinary meaning in the technical field, unless explicitly defined otherwise herein. All references to “a/an/the [element, device, component, means, step, etc.]” are to be interpreted openly as referring to at least one instance of said element, device, component, means, step, etc., unless explicitly stated otherwise. The steps of any method disclosed herein do not have to be performed in the exact order disclosed, unless explicitly stated.

BRIEF DESCRIPTION OF THE DRAWINGS

The above, as well as additional objects, features and advantages of the disclosed embodiments, may be better understood through the following illustrative and non-limiting detailed description, with reference to the appended drawings, where the same reference numerals may be used for similar elements.

FIG. 1 is a schematic view of an example storage system.

FIG. 2 is an example schematic block diagram of a number of data items stored in the storage system.

FIG. 3 is a schematic block diagram of an example data item identifier.

FIG. 4 is a schematic block diagram of an example method for retrieving data.

FIG. 5a-c are illustrations of example communications between different entities in the storage system.

FIG. 6 is a schematic block diagram of an example method for storing data.

FIG. 7 is a schematic block diagram of an example method for deleting data.

DETAILED DESCRIPTION

Detailed embodiments of the disclosed methods and systems may be described with reference to the drawings. The present disclosure is related to a distributed data storage sys-

4

tem comprising a plurality of storage nodes. An example structure of the system and the context in which it may be used is outlined in FIG. 1.

A user computer 1 may access, for example via the Internet 3, an application 5 running on a server 7. The user context, as illustrated here, may be therefore a client-server configuration. However, it should be noted that the data storage system to be disclosed may be useful also in other configurations, for example utilizing other communication methods.

In the illustrated case, two applications 5, 9 may run on the server 7. Of course however, any number of applications may be running on server 7. Each application may have an API (Application Programming Interface) 11 which may provide an interface in relation to the distributed data storage system 13 and may support requests, typically write and read requests, from the applications running on the server. Data may be read and written to the storage system using the methods described in detail in U.S. patent application Ser. No. 13/125,524, filed Apr. 21, 2011, which issued as U.S. Pat. No. 8,688,630 on Apr. 1, 2014, the contents of which are hereby incorporated by reference herein. Methods of reading and writing of data therefore may not be further elaborated upon in detail herein. From an application's point of view, reading or writing information from/to the data storage system 13 may appear to be the same as using any other type of storage solution, for instance a file server or a hard drive.

Each API 11 may communicate with storage nodes 15 in the data storage system 13, and the storage nodes may communicate with each other. Alternatively, or additionally, one or more of the storage nodes 15 may include an API 23 for supporting requests as disclosed above. These communications may be based on TCP (Transmission Control Protocol) and UDP (User Datagram Protocol). Other communication protocols may also be utilized.

The components of the distributed data storage system may be the storage nodes 15 and the APIs 11 in the server 7 which may access the storage nodes 15. The present disclosure may be described in relation to methods carried out in the server 7 and in the storage nodes 15. Those methods may be primarily be embodied as combination software/hardware implementations which are executed on the server and the storage nodes, respectively. The operations of the server and/or storage nodes may together determine the operation and the properties of the overall distributed data storage system.

Although in FIG. 1 the server 7 is illustrated as a member of the storage system 13 which is separate from the storage nodes 15 it should be noted that the server 7 may be a storage node including server functionality.

The storage node 15 may typically be embodied by a file server which is provided with a number of functional blocks. The storage node may thus include a storage medium 17, which for example may include a number of internal (e.g. connected via Integrated Drive Electronics (IDE), Serial Advanced Technology Attachment (SATA), and/or the like) or external hard drives (e.g. connected via Universal Serial Bus (USB), Firewire, Bluetooth, and/or the like), optionally configured as a RAID (Redundant Array of Independent Disk) system. Other types of storage media are however conceivable as well.

Each storage node 15 may contain a node list including the IP addresses of all storage nodes in its set or group of storage nodes. The number of storage nodes in a group may vary from a few to hundreds or thousands of storage nodes.

The storage medium 17 may store one or more data items 19, 21 in the form of collection objects 19 or payload data in the form of data files 21. A collection object 19 may include a set of references. A reference may be a reference to one or

5

more data files stored in the storage system, e.g. data files **21**. A reference may also be a reference to another collection object **19** stored in the storage system. A reference may include a pointer (e.g. a memory address) to a storage location of a storage node **15**. A reference may include an identifier of the collection object or data file referred to.

As will be disclosed in more detail below, the collection object **19** may be used for implementing a structured layer in the storage system. Data files **21** referenced in the collection object **19** may in such an implementation represent data files stored in the structure. Additional collection objects **19** referenced in the collection object **19** may in such an implementation represent subdirectories stored in the directory.

A collection object **19** may be embodied as a data object having a predetermined format. The data object may be a special file in the file system of the storage medium **17** in the sense that it may be a binary file to be interpreted by the API. In an example, the data object may be a standard data file in the file system of the storage medium **17**; the data object may e.g. be a plain text file indicating the referenced collection objects **19** and/or data files **21**. A data object may be readable using the same routines of the file system as the data files **21**.

FIG. **2** schematically illustrates a collection object **19a** according to one embodiment. The collection object **19a** may have an associated collection object identifier **20a**. The identifier **20a** may for instance be a Universally Unique Identifier (UUID). The collection object identifier **20a** may be included in a header of the collection object **19a**. However, the collection object identifier **20a** may be stored in a register maintained at the storage node **15**, for example rather than being included in the collection object **19a**. In an example, the UUID and/or the register maintained at storage node **15** may associate the collection object **19a** with the collection object identifier **20a**, e.g. by pointing out the memory address where the collection object **19a** is to be found. Thus, the collection object **19a** may form a first data item which is identified by a first unique key.

The collection object **19a** may include a field **22a** with an identifier **20b** of another collection object **19b**, e.g. in the form of a string. The collection object **19a** may include a reference to the collection object **19b**. The collection object **19b** may be stored on the same storage node as the collection object **19a** or on another storage node than the collection object **19a**. The storage system may use the identifier **20b** in the field **22a** to locate and access the collection object **19b**. Thus, the collection object **19b** may form a second data item which is identified by a second unique key.

In one embodiment, in order to implement large storage systems spanning over multiple networks, the data item identifiers **20a-d** may include two data elements. With reference to FIG. **3**, the first data element **30** may be a cluster ID **31** that may identify the cluster where the data item (collection object **19a-c** or data file **21a**) is located. The cluster address may be a multicast address **32**. The multicast address **32** may be utilized by the API to send a request for a data item to a specific cluster. The second data element **33** may be a data item ID **34** formed by a unique number **35** that identifies the data item **19a-d** inside the cluster. The unique number **35** may be a number of with a defined length, e.g. 128 bits, or the length may vary. The unique number **35** may include a large number of bits, enabling a large number of data items to be uniquely identified within the cluster. By this arrangement a collection element in one cluster may reference another collection element or data file in another cluster. In other words, the first and second unique key may include a cluster address

6

pointing out a subset of the storage nodes within the system, and a data item identifier identifying a data item within the subset of storage nodes.

Referring back to FIGS. **1** and **2**, the server **7** may for example include a register indicating a storage node **15** storing the collection object (e.g. collection object **19a**) associated with a specific identifier (e.g. identifier **20a**). In another example, the collection object **19a** may be located using the read method disclosed in U.S. patent application Ser. No. 13/125,524. Briefly, according to this read method the server **7** or a storage node **15** may send a multicast message to the plurality of storage nodes **15**. The multicast message may include the identifier **20a** of the desired collection object **19a**. Each storage node **15**, in response to receiving the multicast message, may scan its storage medium **17** for a collection object having said identifier. If found, the storage node **15** may respond and indicate that it stores the object sought-after to the originator of the multicast message. The collection object **19a** may then be accessed by means of a unicast request sent to a responding storage node **15** storing the collection object **19a**.

According to the present embodiment, multicast communication may be used to communicate simultaneously with a plurality of storage nodes. By a multicast or IP multicast is here meant a point-to-multipoint communication which may be accomplished by sending a message to an IP address which may be reserved for multicast applications. For example, a message, for example a request, may be sent to such an IP address (e.g. 244.0.0.1), and a number of recipient servers may be registered as subscribers to that IP address. Each of the recipient servers may have its own IP address. When a switch in the network receives the message directed to 244.0.0.1, the switch may forward the message to the IP addresses of each server registered as a subscriber.

In principle, a single server may be registered as a subscriber to a multicast address, in which case a point-to-point, communication may be achieved. However, in the context of this disclosure, such a communication may nevertheless be considered a multicast communication since a multicast scheme is employed.

According to the present embodiment, unicast communication may refer to a communication with a single recipient. A unicast communication may be initiated by a party of the network and may be directed to a single specific recipient.

In addition to the collection object **19a**, the collection object **19b** may include a field **22b** with an identifier **20c** of a third collection object **19c**. The collection object **19c** may include a field **22c** with an identifier **20d** of a data file **21a**. In other words, anyone of the collection objects **19a-c** (or, for example, each of the collection objects **19a-c**) may represent a second data item including a reference to third data item, and the data file **21a** may represent a second data item including payload data, for example an image.

By appointing the collection object **19a** as a root collection object, the collection object **19a** may represent a root directory **19a** of the storage system. Analogously the collection object **19b** may represent a subdirectory **19b** of the root directory **19a**. The collection object **19c** may represent a subdirectory of the subdirectory **19b**. The data file **21a** may represent a data file stored in the subdirectory **19c**. The collection objects **19a-c** may thus define a hierarchical storage structure. The structure may be referred to as a directory tree.

With reference to FIGS. **4** and **5a-c**, a method for parsing a directory structure in order to access a file **19**, **21** stored in a storage node **15** may be disclosed.

The starting point of the directory structure may be a predefined root key. For example, any of the storage nodes **15**

7

may include a root key. This key may be stored outside of the storage cluster and may be used to identify the first data item (e.g. collection object **19a**) in the directory structure. A storage cluster may have multiple root keys that enable the user to have multiple individual directory structures stored within the same storage cluster. The directory structures may span over several storage clusters. The root key may be stored together with external information describing the directory structure stored within the cluster.

In block **40**, the server **7** may receive the root key, which may identify the first data item **19**, **21** and may pass the unique identifier to identify the file within the storage system to the API **11**. In an example, the API **23** may be implemented in a storage node **15**, wherein the root key may be received in the storage node **15** rather than in the server **7**.

In block **41**, the API **11** in the server **7** may multicast a request for the data item (e.g. the collection object **19a**) identified by the root key to storage nodes **15a-e** in the storage system, or to a subset of the nodes. For example, the multicast message may be sent to a specific cluster, for example using the data item identifier configuration disclosed in relation to FIG. **3**. According to one embodiment the data item (e.g. the collection object **19a**) identified by the root key may be a special data item in the sense that it may include additional metadata which may be used by the system. Examples of such data may be information regarding access permissions to the items in the directory structure, information where to store certain data items (e.g. on a storage node with quick access such as an Solid State Drive (SSD)), and the like.

In block **42**, storage node **15a-e**, in response to receiving the multicast message, may scan their respective storage mediums **17** in an attempt to locate the data item identified by the data item ID **34** in the root key.

For purposes of illustration, it may be assumed in this example that nodes **15b** and **15e** locate the data item identified by the data item ID **34**. In block **43**, the nodes **15b**, **15e** that find the data item may reply with information about what other nodes **15b**, **15d**, **15e** may contain the data item and the current execution load (e.g. how busy the nodes are, how many request the nodes received, how much free space is on the node, etc.) in the node **15b**, **15e**. The requested data item may be stored in a plurality of storage nodes **15b**, **15d**, **15e**, wherein the API may collect the information received from the nodes **15b**, **15d**, **15e** and may wait until it has received answers from more than 50% of the listed storage nodes **15b**, **15d**, **15e** that contains the data item before it may make a decision on which one to select for the retrieval of the data item. The decision may be based on which node that has the lowest execution load.

In block **44** the API **11** may send a unicast request for the specific file to the chosen storage node. In this example, for purposes of illustration it may be assumed store node **15b** is chosen. API **11** may retrieve the data item from the storage node **15b**. The API **11** may maintain a list of all storage nodes **15b**, **15d**, **15e** that store copies of the file in the event of a read or communication failure with the selected node **15b**. If an error occurs the API **11** may transparently select the next best node in the list and continues the read operation.

In block **45** the API may interpret the content of the retrieved data item. If the directory structure comprises additional levels the retrieved data item may be a collection object **19b**. If so, the API **11** may read the field **22b** which may include an identifier **20b** referring to another collection object **19c** in the directory structure. For example, the API may retrieve the unique key, i.e. the identifier **20b**, identifying the second data item, e.g. the collection object **19b**, from the received first data item, e.g. the collection object **19a**. The

8

process may then return to block **41** and may continue parsing the directory structure. Thus, both the first and second request may be sent from an application programming interface, API. The process may continue until the last object in the directory structure has been identified and retrieved, e.g. the data file **21a** whereupon the process may end at **46**. In another example, the API **11** may send an update request to the identified object, e.g. a command to alter or concatenate data in the data item corresponding to the object in the directory structure.

As an example, it may be that the data file **21** is located in the root of the directory structure. In such a case the process may be looped a single time, since the first retrieved collection object **19a** may contain a reference to the data file **21a**. It is emphasized that the retrieved collection object in addition to including the reference to the data file **21a** may also include references to other data items, such as collection object **19b**.

Thus, according to the above, a method may implemented in a data storage system including data storage nodes interconnected by means of a communications network for accessing file. The method may include sending a request for a first data item **19**, **21**, (e.g. collection object **19a**), to a plurality of storage nodes **15a-e**. The first data item may include a reference to a second data item (e.g. data file **21a** or collection object **19b**), stored in the storage system. The method may include receiving the first data item from at least one storage node **15b**, and sending a request for the second data item to the plurality of storage nodes **15a-e** based on the reference included in the first data item.

As an illustrative example, with reference to FIG. **2**, the API may recursively read and interpret referenced envelopes to resolve a path in a directory structure. For example, the API may identify an unstructured key that represents a file in the structured path. For example, a user accessing the storage system may want to resolve the path: “/Documents/Sample_Pictures/Blue_Hills.jpg”.

In FIG. **2**, collection object **19a** may represent the root key “I” (identified by the unique key **20a**) and the identifier **22a** may include a reference to collection object **19b** representing the folder “Documents/” (identified by the unique key **20b**). The identifier **22b** in the collection object **19b** may include a reference to collection object **19c** representing the folder “Sample_Pictures/”. Finally, the identifier **22c** in the collection object **19c** may include a reference to the data file **21a** comprising the payload data for the file “Blue_Hills.jpg”. Thus, by recursively reading the references in the collection objects a virtual file structure may be created in an unstructured storage system.

With reference to FIGS. **6** and **5a-c**, a method for parsing a directory structure in order to create a file **19**, **21** in a storage node **15** is disclosed.

similarly to the system disclosed in FIG. **4**, the starting point of the directory structure is a predefined root key. The root key may be an arbitrary key, and there may be many root keys throughout the system. This key may be stored outside of the storage cluster and may be used to identify the first data item (e.g. collection object **19a**) in the directory structure.

In block **60**, the server **7** may receive the root key, and may pass the unique identifier to identify the file within the storage system to the API.

In block **61**, the API **11** may resolve the path to the desired data item according to the method above.

In block **63**, the API **11** in the server **7** may multicast a request for storing the data item (e.g. the collection object **19c**) including the identifier to all storage nodes **15a-e** in the storage system, or to a subset of the nodes, e.g. within a

specific cluster, for example using the data item identifier configuration disclosed in relation to FIG. 3.

In block 63, storage nodes 15a-e, in response to receiving the multicast message, may verify that the data item ID 34 is not already in use.

In block 64, a storage node which 15a-e that fails to find an existing file with that specific identifier may reply with an acknowledgment that may indicate: free storage space on the storage node, an indication of the age of the hardware that the storage node is running on, current CPU load and/or the geographical position of the storage node 15a-e in the form of latitude, longitude and altitude or the like.

In block 65 the API 11 may select three storage nodes (e.g. storage nodes 15a, 15b and 15e) based on the data returned from storage nodes that responded to the multicast request. When the three most suitable nodes have been selected the API 11 may send a request to the three nodes simultaneously to store the data item. If an error occurs during the transfer of the data item to one of the selected nodes 15a, 15b, 15e, the operation continues, for example as long as more than 50% of the selected nodes are operational.

In block 66 the identifier field 22b in the data item one level higher up in the directory structure (e.g. the first data item—the collection object 19b) may be updated with a reference to the stored data item (e.g. the collection object 19c), either by retrieving the first data item according to the read method according to the above or by accessing the first data item directly, for example if the server has cached the identifier of the first data item.

In order to increase the data integrity in the system the method above may be supplemented with the act of, prior to storing the data item, retrieving the first data item in case the communication with all storage nodes should be lost after the data item has been stored but before the first data item is updated. By this procedure the API may resume the update procedure once the communication with the storage nodes is resumed.

Thus, according to the above, a method may be implemented in various devices within a data storage system including data storage nodes interconnected by means of a communications network. The method may include storing a first data item in at least one storage node, and updating a second data item, stored in at least one storage node, by adding a reference to the first data item in the second data item.

With reference to FIGS. 7 and 5a-c, a method for parsing a directory structure in order to delete a file 19, 21 in a storage node 15 is disclosed.

In similarity to the disclosure in relation to FIG. 4, the starting point of the directory structure may be a predefined but arbitrary root key. This key may be stored outside of the storage cluster and may be used to identify the first data item (e.g. collection object 19a) in the directory structure.

In block 70, the server 7 may receive the root key and may pass the unique identifier to identify the file within the storage system to the API.

In block 71, the API 11 may resolve the path to the desired data item according to the method above.

In block 72, the API 11 in the server 7 may multicast a query regarding the location of the data item (e.g. the collection object 19c) including the identifier to storage nodes 15a-e in the storage system, or to a subset of the nodes, e.g. within a specific cluster, for example using the data item identifier configuration disclosed in relation to FIG. 3.

In block 73, storage nodes 15a-e, in response to receiving the multicast message, may scan their respective storage medium 17 to locate data item identified by the data item ID 34.

In block 74, the nodes that locate the data item may reply with information regarding other nodes that may store the data item and the current execution load in the node. The requested data item may be stored in a plurality of storage nodes. The API may collect the information received from the nodes and may wait until it has received answers from more than 50% of the listed storage nodes that contains the data item before making a decision on which nodes to select for the deletion of the data item.

In block 75 the API 11 may send a unicast request to delete the specific file (e.g. the collection object 19c) to the chosen storage nodes.

In block 76 the identifier field 22b in a data item one level higher in the directory structure (e.g. the collection object 19b) may be updated by deleting the reference to the deleted data item (e.g. the collection object 19c). The update may occur by retrieving the first data item according to the read method described above and/or by accessing the first data item directly, for example if the server cached the identifier of the first data item. In the case where the data item to be deleted is located a number of levels down in the directory structure, the delete operation may be expressed as the method disclosed in relation to FIG. 4 with the addition of i) deleting the first data item, and ii) updating the second data item by deleting the reference to the first data item.

Thus, a data deletion method may be implemented in a data storage system including data storage nodes interconnected by means of a communications network. The method may include deleting a first data item stored in at least one storage node. The method may also include updating a second data item, stored in at least one storage node, by deleting a reference to the first data item in the second data item.

The collection objects 19 may be handled and maintained in manner similar to data files. This may allow the data to be stored in a flat storage structure, e.g. without any subdirectories or within a single directory. A virtual hierarchical storage structure may be created by adding collection objects 19 including references to other collection objects 19 and/or data files 21. It even allows the same data to be organized in several different virtual hierarchical storage structures by using different sets of collection objects 19.

For data security reasons, some or all information stored in the storage system (e.g. the collection objects 19 and/or the data files 21) may be stored redundantly in the storage system. The collection objects 19a-c and the data file 21a may be stored at two or more storage nodes 15. Each instance of a collection object or data file may be associated with the same identifier. In such a case, the above-described read method may result in a response from each storage node storing the collection object. A redundantly stored collection object may thus be retrieved from either one of or all of the storage nodes storing the collection object.

Several embodiments which demonstrate the disclosed method and system have been described. However, as is readily appreciated by a person skilled in the art, other embodiments in addition to the ones disclosed above are equally possible that are in accordance with the methods and products described herein. The foregoing examples are not meant to be limiting, and the scope of the protection is to be defined by the appended patent claims.

11

What is claimed:

1. A method for implementing a distributed directory structure in a data storage system, the method comprising:

identifying a first data item in a data storage system, wherein the first data item comprises a first collection object, the first collection object comprises one or more references to one or more other data items in the data storage system, the one or more references comprise a reference to a second data item stored in the data storage system, and the first collection object corresponds to a first subdirectory in a distributed directory system; determining a first multicast address to use for requesting the first collection object; sending a first request for the first collection object using the first multicast address, the first request including a unique identifier for the first collection object; receiving at least one copy of the first collection object based on sending the first request using the first multicast address; determining a second multicast address used to request the second data item in a cluster of the data storage system, wherein the reference to the second data item included in the first collection object comprises the second multicast address, the second data item comprises a second collection object, the second collection object corresponds to a second subdirectory, and the first subdirectory is a parent directory of the second subdirectory; sending a second request for the second data item to the second multicast address included in the reference to the second data item included in the first collection object; and receiving the second data item via a unicast message based on the second request.

2. The method as in claim 1, wherein multiple copies of the second data item are received based on the second request, each of the multiple copies being received from a different data storage node of the cluster.

3. The method as in claim 1, wherein the first data item and the second data item are stored in different clusters of data storage nodes.

4. The method as in claim 1, wherein the second collection object comprises at least one reference to at least one other data item in the data storage system, and the at least one reference comprises a reference to a third data item stored in the data storage system.

5. The method as in claim 4, wherein each of the first data item, the second data item, and the third data item are stored on different data storage nodes in the data storage system.

6. The method as in claim 1, wherein a collection object is a type of file utilized to implement the distributed directory structure.

7. The method as in claim 6, wherein collection objects in the data storage system each comprise at least one of a reference to another collection object or a reference to a data file.

8. The method as in claim 7, wherein the collection objects in the data storage system each represent a folder or subfolder of the distributed directory structure.

9. The method as in claim 1, wherein the first collection object also includes a reference to a data file, and the data file comprises payload data.

10. A method for locating a data file within a distributed directory system, the method comprising:

determining a path for accessing a first data item, wherein the first data item comprises a data file, the path for accessing the first data item includes one or more subdirectories, and each of the one or more subdirectories is

12

represented in the distributed directory system using a corresponding collection object;

determining a first multicast address to use for requesting a first collection object that represents a first subdirectory in the path;

sending a first request for the first collection object using the first multicast address, the first request including a unique identifier for the first collection object;

receiving at least one copy of the first collection object based on sending the first request using the first multicast address, wherein the first collection object comprises a second multicast address for a second collection object representing a second subdirectory, the first subdirectory being a parent directory of the second subdirectory;

sending a second request for the second object using the second multicast address, wherein sending multicast requests for subsequent collection objects representing further subdirectories are continued until at least a subdirectory that includes the first data items is located; and receiving the first data item via a unicast message.

11. The method as in claim 10, wherein a copy of each collection object is stored on multiple data storage nodes.

12. The method as in claim 11, further comprising determining which copy of a identified collection object is to be requested based at least in part on load information for the data storage nodes that store the identified collection object.

13. The method as in claim 10, further comprising using a root key to identify a collection object representing a root directory for the path, and requesting the collection object representing the root directory based on the root key.

14. A server implemented at least in part using hardware, the server comprising a processor configured to:

identify a first data item in a data storage system, wherein the first data item comprises a first collection object, the first collection object comprises one or more references to one or more other data items in the data storage system, the one or more references comprise a reference to a second data item stored in the data storage system, and the first collection object corresponds to a first subdirectory in a distributed directory system; determine a first multicast address to use for requesting the first collection object;

send a first request for the first collection object using the first multicast address, the first request including a unique identifier for the first collection object;

receive at least one copy of the first collection object based on sending the first request using the first multicast address;

determine a second multicast address used to request the second data item in a cluster of the data storage system, wherein the reference to the second data item included in the first collection object comprises the second multicast address, the second data item comprises a second collection object, the second collection object corresponds to a second subdirectory, and the first subdirectory is a parent directory of the second subdirectory;

send a second request for the second data item to the second multicast address included in the reference to the second data item included in the first collection object; and receive the second data item via a unicast message based on the second request.

15. A data storage node comprising a processor configured to:

identify a first data item in a data storage system, wherein the first data item comprises a first collection object, the first collection object comprises one or more references

13

to one or more other data items in the data storage system, the one or more references comprise a reference to a second data item stored in the data storage system, and the first collection object corresponds to a first subdirectory in a distributed directory system;

determine a first multicast address to use for requesting the first collection object;

send a first request for the first collection object using the first multicast address, the first request including a unique identifier for the first collection object;

receive at least one copy of the first collection object based on sending the first request using the first multicast address;

determine a second multicast address used to request the second data item in a cluster of the data storage system, wherein the reference to the second data item included in the first collection object comprises the second multicast address, the second data item comprises a second collection object, the second collection object corresponds to a second subdirectory, and the first subdirectory is a parent directory of the second subdirectory;

14

send a second request for the second data item to the second multicast address included in the reference to the second data item included in the first collection object; and

receive the second data item via a unicast message based on the second request.

16. The data storage node as in claim **15**, wherein the processor is configured to add a new data file to the second subdirectory corresponding to the second collection object by adding a reference to the new data file to the second collection object.

17. The data storage node as in claim **15**, wherein the processor is configured to add a new subdirectory to the first subdirectory corresponding to the first collection object by adding a reference to a third collection object to the first collection object, the third collection object corresponding to the new subdirectory.

18. The data storage node as in claim **15**, wherein the processor is configured to delete the second subdirectory corresponding to the second collection object by removing the reference to the second collection object from the first collection object.

* * * * *